DOCUMENT RESUME

ED 194 636 TM 800 768

AUTHOR Engelhard, George, Jr.
TITLE An Introduction to Rasch Measurement and Its Application to Test Equating in the Comprehensive Assessment Program.
PUB DATE May 80
NOTE 27p.: Paper presented at the Annual Meeting of the Northern Illinois Association for Educational Research, Evaluation and Development (4th, Bloomingdale, IL, May 9, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests: Difficulty Level: Elementary Secondary Education: *Equated Scores: *Latent Trait Theory: Mathematical Models: Reading Tests: Student Evaluation: Test Items: *Test Theory
IDENTIFIERS *Comprehensive Assessment Program: *Rasch Model: Unidimensional Scaling: Vertical Equating

ABSTRACT

The Rasch model is described as a latent trait model which meets the five criteria that characterize reasonable and objective measurements of an individual's ability independent of the test items used. The criteria are: (1) calibration of test items must be independent of particular norming groups: (2) measurement of individuals must be independent of particular items used for measuring: (3) test items must measure a single underlying trait or ability: (4) a more able individual must have a better chance of success with an item than a less able individual: and (5) an individual must have a better chance of success on an easy item than on a difficult item. To illustrate the use of the Rasch model for vertical equating, equal interval scales (EIS) were developed and applied to the Scott, Foresman Comprehensive Assessment Program (CAP), a coordinated series of tests and measures for evaluating students' educational growth. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194636

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Englehard

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

AN INTRODUCTION TO RASCH MEASUREMENT

And

ITS APPLICATION TO TEST EQUATING

In The

COMPREHENSIVE ASSESSMENT PROGRAM

George Engelhard, Jr.
University of Chicago

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 800768

Paper presented at the Northern Illinois Association for Educational Research, Evaluation and Development meeting in Bloomingdale, Illinois, May, 1980.

Abstract

The purpose of this paper is to provide a basic introduction to the Rasch model and to illustrate its use for equating psychological and educational tests. The data used for the equating example was taken from a set of standardized reading tests which are a part of the Achievement Series of the Comprehensive Assessment Program (Scott, Foresman and Company, 1980.)

AN INTRODUCTION TO RASCH MEASUREMENT AND ITS APPLICATION TO TEST EQUATING

INTRODUCTION

One of the major problems encountered in educational measurement is the equating of person measurements obtained on different tests. This problem occurs whenever the variable of interest is represented by a range of item difficulties which go beyond the ability of any one group of individuals to attempt. For example, as educators we may be interested in tracing an individual's growth over the elementary school years. Any single test that we might use would be much too difficult for first graders and much too simple for eighth graders. If we use multiple tests, composed of items whose difficulties are appropriate for each person's level of ability, then we are faced with the problem of determining the equivalence or comparability of measures obtained from several different measuring instruments. A solution to the equating problem can be found, if we can create several tests composed of items calibrated onto a single scale which represents a unidimensional construct (e.g., reading ability) and spans the time period over which we wish to measure growth. In order to accomplish this goal, we need a method for equating tests and linking items together. These linked items can be used to represent the latent construct or variable of interest on which we wish to measure an individual's growth and change.

This problem was recognized by Thorndike in the early 1920s.

> With the development of group tests and tests for use with higher levels of intelligence, it is becoming more and more necessary to transmute a score obtained with one test into the score that is equivalent to it in some other test.
>
> (Thorndike, 1922, p. 29)

Various methods have been proposed as solutions to the equating problem. Thorndike "transmuted" scores using his probable error method of scaling (Thorndike, 1922; Trabue, 1916). Thurstone in a series of articles in the 1920s described his absolute scaling method which he proposed as a solution to the equating problem (Thurstone, 1925, 1927, 1928). More recently, latent trait measurement theory has been recommended as a source of solutions to the "intractable" problem of equating (Lord, 1977; Marco, 1977; Rasch, 1960; Wright, 1967; Wright and Stone, 1979).

In his extensive discussion of equating, Angoff (1971) listed what he considered two "restrictions" or what may be better thought of as reasonable assumptions and conditions necessary in order to equate tests. These conditions are:

1. that the two instruments (tests, items) in question be measures of the same characteristics in the same sense that degrees Fahrenheit and centigrade, for example, are both units of temperature, inches and centimeters are both units of length, etc. (unidimensionality condition)

2. that, in order to be truly a transformation of systems of units, the conversion must be unique, except for the random error associated with the unreliability of the data, and the method used for determining the transformation; the resulting conversion should be independent of the individuals from whom the data were drawn to develop the conversion and should be freely applicable to all situations. (sample-free condition)

The first condition for acceptable equating involves the unidimensionality of the measures to be equated, while the second condition implies a sample-free procedure for equating. Both of these conditions are necessary in order to realize the advantages of equated tests. All of the previously proposed

procedures can meet or approximate the first condition of unidimensionality. None of the methods proposed prior to the development of latent trait measurement theory meet the sample-free condition and only one set of latent trait models--Rasch measurement models--offers reasonable solutions to the problem of sample-free equating (Engelhard, 1980).

The purpose of this paper is to provide a basic introduction to the Rasch model and to illustrate its use for equating educational tests. The data used for the equating example was taken from a preliminary set of reading tests which are part of the Comprehensive Assessment Program (Scott, Foresman and Company, 1980).

INTRODUCTION TO THE RASCH MODEL

During the 1950s, Georg Rasch conducted the basic psychometric work which led to the publication in 1960 of his book, Probabilistic Models for Some Intelligence and Attainment Tests. The ideas and methods presented in this book represent some of the most innovative and useful work in psychometrics, since Thurstone's work in the 1920s. In fact, Rasch's work represents an almost totally new approach to psychometrics.

In traditional or classical psychometrics, the properties of a test are defined in terms of variations within some specified population of people. As a consequence, the properties of the test, e.g., the reliability coefficient, are not specific to the test itself, but will vary depending on the population chosen. Similarly, the measurement of a person on the variable of interest will depend on which items are used. In traditional approaches to person measurement, the ability estimate depends not only on which items are used, but also on the group of people with which the person

is compared. As Wright has pointed out,

> if all of a specified set of items have been tried by a child you wish to measure, then you can obtain his percentile position among whatever groups of children were used to standardize the test. But how do you interpret this measure beyond the confines of that set of items and those groups of children? Change the children and you have a new yardstick.
>
> Change the items and you have a new yardstick again. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardsticks?
>
> (Wright, 1967, p. 86)

The use of Rasch measurement models provides a reasonable solution to the problem of "rubber yardsticks" by providing estimates for intrinsic properties of tests and items which are independent of the group that happens to be used to calibrate the items. This is called person-free item calibration. It also yields estimates of a person's ability which are independent of the test items used. This leads to the possibility of item-free person measurement. Of course this does not mean that we can measure people without items, but it does mean that once items are calibrated through the use of the Rasch model and assigned a position on the latent variable of interest, then any set of items can be used to obtain an estimate of a person's ability. These two consequences--person-free item calibration and item-free person measurement--are necessary in order to have objective measures.

In order to obtain reasonable and objective measurement, the measurement model utilized must satisfy at least the following five conditions. These conditions are that:

1. the calibration of test items must be independent of the particular individuals used for the calibration.

2. the measurement of individuals must be independent of the particular items that happen to be used for the measuring.

3. the test items must be measuring a single underlying trait or ability.

4. a more able individual must always have a better chance of success on an item than a less able individual.

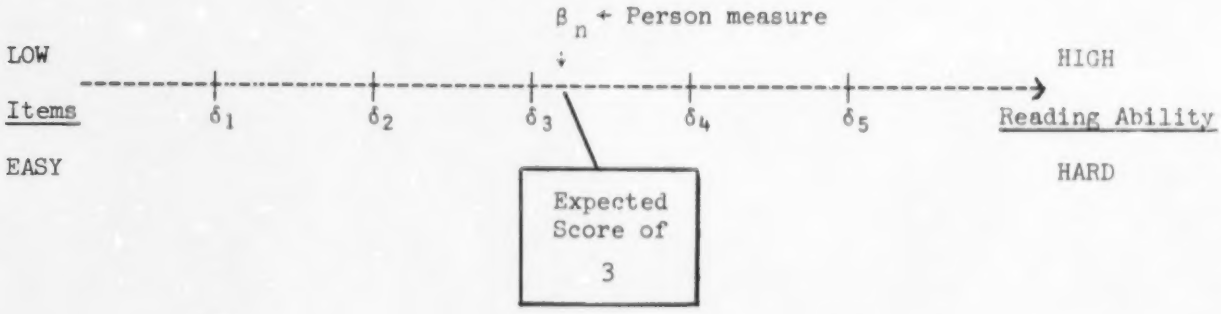
5. any individual must have a better chance of success on an easy item than a more difficult item.

The Rasch model is a latent trait model that has been proposed for person measurement that meets these five conditions. Basically, latent trait models are idea or inventions that attempt to specify what happens when a person tries an item. (See Hambleton and Cook (1977) for a general introduction to latent trait models.)

Of all the latent trait models, Rasch measurement models have the fewest ingredients, one ability parameter, β_n, for each person n and one difficulty parameter, δ_i, for each item. These parameters represent the position or location of persons and items on the latent variable. For example, if the latent variable is reading ability, we develop and choose a set of items to represent this variable. These items are then given to a group of people and their locations are determined through the application of

the Rasch model. The locations of the people on the latent variable reading ability, are given by the ability estimates, while the locations of the items are given by the difficulty estimates. This is illustrated in Diagram 1.

Diagram 1. Defining a variable.

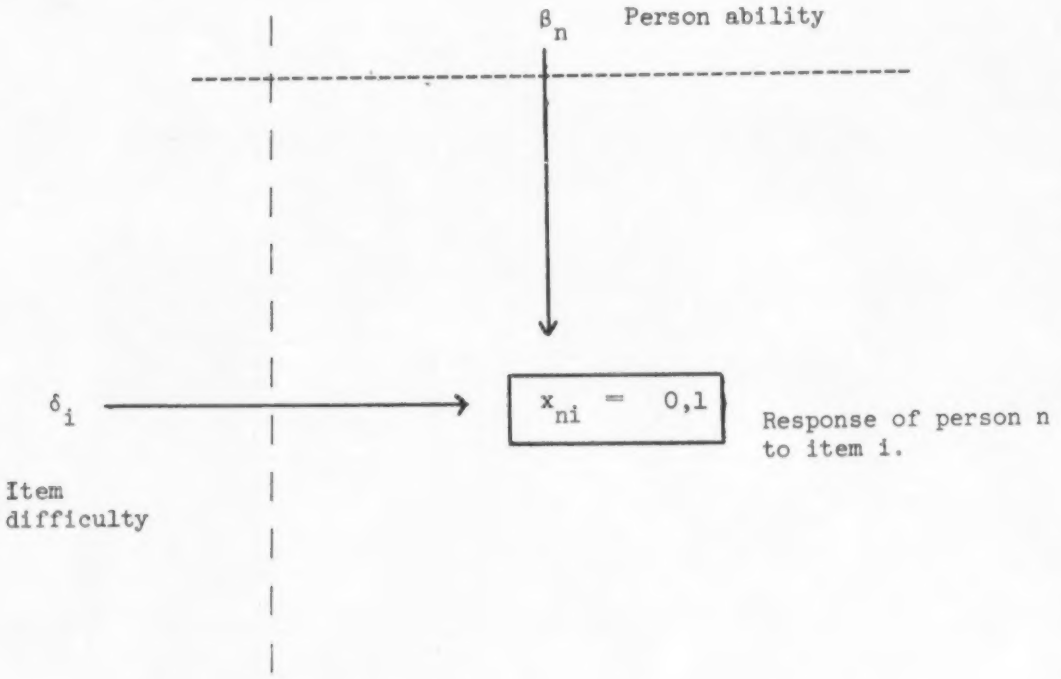


In Diagram 1, the line represents the latent variable called reading ability. Five items have been chosen to represent this construct and their difficulties which locate them on the latent variable are shown below the line (δ_1 to δ_5) The items range from easy on the left to more difficult on the right. Person measurements are shown above the line and in this case there is one person measure. This person correctly answered items 1 to 3 and incorrectly answered items 4 and 5. This person's score would be 3 and this value can be used to locate the person on the latent variable by providing an estimate of reading ability.

The ability parameters and difficulty parameters are combined in order to represent one latent dimension by forming their difference ($\beta_n - \delta_1$). This difference governs the probability of what happens when person n attempts

item i. The basic data which we have in any testing situation is a matrix of 0's and 1's which represent each individual's failure (0) or success (1) on each item. This is illustrated in Diagram 2.

Diagram 2. The essential conditions causing a response.



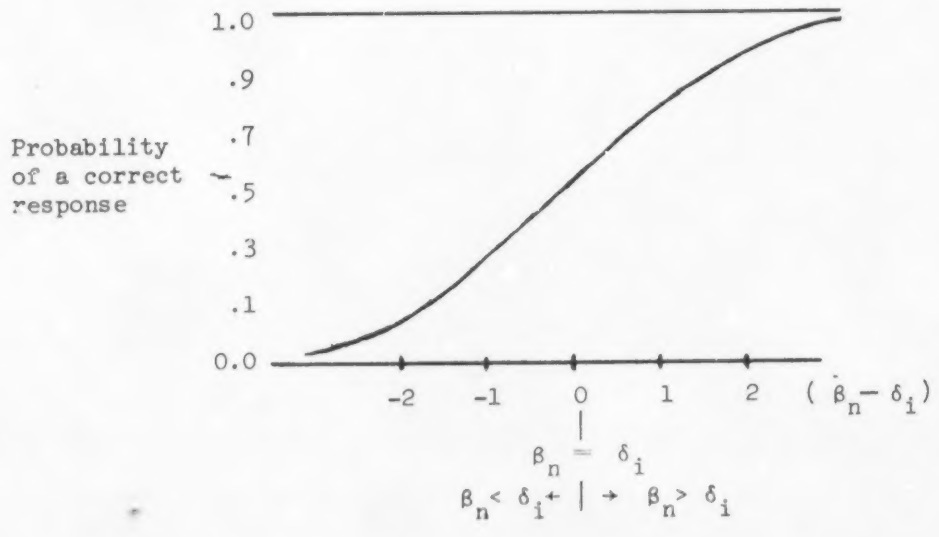
The mathematical model used to express this relationship is shown in Diagram 3.

Diagram 3. Mathematical formulation of the Rasch model with two response categories.

$$\Pr\{x_{ni} = 0,1 \mid \beta_n, \delta_i\} = \frac{\exp x_{ni} (\beta_n - \delta_i)}{[1 + \exp (\beta_n - \delta_i)]}$$

The probability of observing a correct response (1) or an incorrect response (o) for person n on item i is a function of the difference between the person's ability (β_n) and the item's difficulty (δ_i) The relationship represented by the Rasch model between this difference ($\beta_n - \delta_i$) and probability of success on an item can be illustrated with an item characteristic curve or response curve in which the item difficulty remains constant, while person ability varies. (See Diagram 4). If the person's ability equals the difficulty of the item, then the person has a 50% chance of success on that item. In other words, the person can be expected to succeed half the time on this kind of item, and conversely to fail half the time. If the person's ability exceeds the item's difficulty, then a person has a better than 50% chance of success on the item; if the item's difficulty exceeds the person's ability to answer the item correctly, then the person has a less than 50% chance of success.

Diagram 4. Response Curve.



TEST EQUATING WITH THE RASCH MODEL

As pointed out earlier, various procedures have been proposed for test equating, but the only method which meets all the conditions necessary for objective equating is the Rasch model. Our goal in test equating is to step beyond the specific items contained in separate tests in order to get information on the latent trait or unobservable variable which is of interest. Since no individual can handle the full range of difficulties, it is necessary to translate the measures obtained on different tests into one common metric on a unidimensional scale that represents the latent variable. For example, suppose we are interested in measuring the change and growth in reading ability of students from grade 3 to grade 4. If the students were given exactly the same test, many of the students in the beginning of grade 3 would experience frustration when attempting items appropriate for them at the end of grade 4; these items would be obviously too difficult and thus inappropriate for a grade 3 student. Conversely, when these students are in grade 4, they might become bored with items appropriate for grade 3 students and now obviously too easy. In addition to these extraneous influences on the measuring situation, there are problems, such as memory effects, that arise when children are retested using the same tests. The well known fact that ability estimates are most accurate when they are based on items of appropriate difficulty for the student must also be considered. One approach is to link several tests together with a subset of carefully chosen items, so that the students are taking tests which are appropriate for their ability which will minimize extraneous in-

fluences on the measuring situation and provide more accurate estimates of an individual's ability or location on the latent trait. This link of common items can be displayed as shown in Diagram 5.

Diagram 5. Common item link.

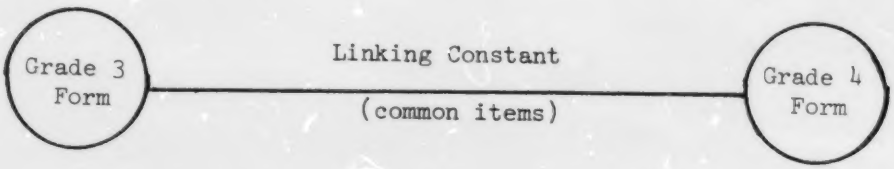
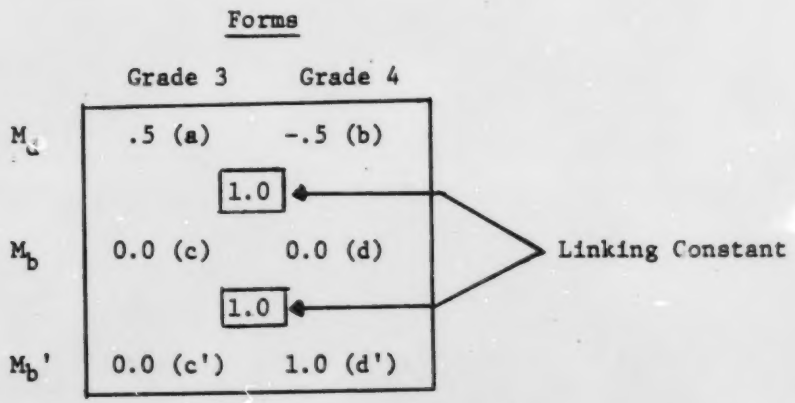


Figure 2 illustrates this type of display with the actual linking constants in position for several forms which measure reading ability over a 6 year period.

The basic logic behind the linking of tests through common items can be illustrated using the following table, which is based on hypothetical data.



Suppose students in grades 3 and 4 each have taken separate test with 10 common items. The average difficulty estimates (M_d) of these 10 items for each group is shown in cells (a) and (b). In order to compute these estimates, separate calibrations are conducted on each test using the Rasch model. (See Wright and Mead (1976) for a description of the calibration procedure and a computer program that can be used to obtain these estimates.) The next step is to take the two independent difficulty estimates for the 10 common items and compute the two mean difficulties which were obtained through the separate calibrations using each grade. The average ability estimates (M_b) for each grade are centered at zero in the usual way (Wright and Stone, 1979). Since the items are the same, they should represent the same point and location on the latent trait scale. In other words, the difficulty estimates for the common items should ideally be the same whether they are determined with grade 3 students or grade 4 students. In order to approximate this equality, we take the average difference between the independent difficulty estimates as a linking constant (or translation constant) that can be used to bring the difficulty estimates together. Because of the assumptions and properties of our measurement model, the relationship between the (a) and (b) cells should also hold for the (c) and (d) cells. In order to maintain the equality of these relationships, we simply add the linking constant, 1.0, to each of the estimated abilities of the grade 4 students. The addition of this linking constant yields the revised estimates of mean abilities (M_b') which represents the location of the mean ability on one unidimensional scale that

spans grades 3 and 4. The extension of this logic and procedure to several tests over a longer time period is straight forward.

METHOD

Item response data from a national sample of greater than 70,000 students were obtained from Scott, Foresman and Company. These data were used for the standardization and calibration of the Comprehensive Assessment Program (C. A. P.). The Comprehensive Assessment Program is a coordinated series of tests and measures for evaluating students' educational growth. In order to accomplish the goal of evaluating educational growth in achievement, equal interval scales (EIS scales) were developed using the Rasch model for the four substantive areas of reading, mathematics, language and study skills.

In order to illustrate the application of the Rasch model to the problem of vertical equating, Forms 3A and 4B from the elementary Achievement Series was used. There were 14 common items and the independent estimates of the difficulties (along with their standard errors in parentheses) are given in columns one and two in Table 1. The next step is to compute the average difference in these difficulty estimates which is shown in column three. The mean of this difference is 1.22 (standard deviation of .37) which provides the preliminary estimate of the linking constant.

The next task was to assess the fit of the items to the link. A two-step procedure was employed to accomplish this. First, the difficulties were plotted and approximate 95% confidence intervals developed. According to the Rasch model, the plot should define a 45° line (slope of 1), so that a constant (or mean difference) is the only adjustment required. Figure 1 shows the bivariate plot of the difficulty estimates for the 14 common items. The items represented by the black circles are vocabulary items and there is

some question about their contribution to the quality of the link. The second step was to examine the residuals. This residual analysis is summarized in columns four through six in Table 1. The standardized residuals verify the conclusions drawn from the plot of the difficulties that the vocabulary items do not fit as well as the reading comprehension items. These standardized residuals are partially inflated due to the very small standard errors of the item difficulties. These standard errors are small because of the large sample size which provides extremely precise difficulty estimates, but tend to make the statistical tests of fit overly sensitive to outliers. A decision rule using the root mean square, which is more robust and less sensitive to outliers, could be developed. In practical situations, the decision rule to reject linking items becomes a substantive issue rather than a statistical one. In the present example, the four largest standardized residuals were associated with vocabulary items. The decision was made to delete these items from the link and the computation of the revised linking constant of 1.019 (rounded to 1.02) is given in Table 2.

The final task in developing an equal interval scale based on the Rasch model is to take the linking constants and add them to the ability estimates obtained on each form which serves to translate the raw scores on each form into the same metric on the latent variable of reading ability. Table 3 gives the adjusted ability estimates in logits for the corresponding raw scores on each form. Starting with form 2B, the mean ability estimates are centered at zero (mean = -.006). In order to link scores on form 3A and make them equivalent to ability estimates derived from

form 2B, the linking constant of 1.03 is added to the initial ability estimates and centered at 1.03 (mean = 1.03). In the last column of Table 3, form 4B is linked to the other two tests by adding the linking constant of 2.05 which is the sum of the link between forms 2B and 3A (1.03), and the link between form 3A and 4B (1.02). It should be pointed out that the equal interval scale is centered on form 2B, so that the linking constants accumulate as we move across the forms. Once the forms are equated and a table like Table 3 is constructed, it is very easy to obtain equivalent ability estimates independent of the forms used to provide the estimates. In other words, if a student's ability in logits was approximately 1.00, we would expect the raw scores of 70 on form 2B, 50 on form 3A, and 30 on form 4B.

DISCUSSION

The Rasch model provides a clear and practical method for equating educational and psychological tests. It is the only equating method based on latent trait measurement that can meet the second condition necessary in order to equate tests; namely, the sample-free condition. The other latent trait models, by including parameters for item discrimination and guessing, provide sample-dependent item and person statistics. The specific objectivity, which is provided by Rasch measurement models, yields the possibility of objective equating. Objective measurement and equating are necessary in order to measure student growth in achievement and in order to measure educational development.

SUMMARY

The first section of this paper provided an introduction to the Rasch model. In the second section a detailed illustration of the application of the Rasch model to the problem of vertical equating was developed.

Table 1. Analysis of item links for equating Form 3A and Form 4B (Reading).

Item Name	Form 3A (d_a)	Form 4B (d_b)	Difference ($D = d_a - d_b$)	Residual Difference (D - 1.22)	S.E. Residual (S_D)	Standardized Residual $z = (D - 1.22)/ S_D$
1	.788(.039)	-.183(.064)	.971	-.249	.075	-3.32
2	.489(.040)	-.666(.068)	1.155	-.065	.079	-.82
3	.419(.040)	-.457(.066)	.876	-.344	.077	-4.47
4	1.119(.040)	-.207(.064)	1.326	.106	.075	1.41
5	.711(.039)	-.215(.064)	.926	-.294	.075	-3.92
6	-.295(.042)	-1.217(.074)	.922	-.298	.085	-3.51
7	-.335(.042)	-1.227(.074)	.892	-.328	.085	-3.86
8	.396(.040)	-.508(.066)	.904	-.316	.077	-4.10
9	-.032(.041)	-1.260(.074)	1.228	.008	.085	.09
10	-.032(.041)	-1.014(.071)	.982	-.238	.082	-2.90
11	1.115(.040)	-.798(.068)	1.913	.693	.079	8.77
12	.268(.040)	-1.396(.076)	1.664	.444	.086	5.16
13	.481(.040)	-1.310(.075)	1.791	.571	.085	6.72
14	1.287(.040)	-.293(.065)	1.580	.360	.076	4.74
Mean	.455	-.767	1.22	.000		-.001
S.D.	.514	.461	.37	.37		4.580

Table 2. Analysis of items retained and used to compute linking constant between form 3A and form 4B.

Item Name	Difficulty Difference	Residual Difference	S.E. Residual	Standardized Residual
1	.971	-.047	.075	-.63
2	1.155	.137	.079	1.73
3	.876	-.142	.077	-1.84
4	1.326	.308	.075	4.12
5	.926	-.092	.075	-1.23
6	.922	-.096	.085	-1.13
7	.892	-.126	.085	-1.48
8	.904	-.114	.077	-1.48
9	1.228	.210	.085	2.47
10	.982	-.036	.082	-.44
Mean	1.018	-.000		.01
S.D.	.16	.16		2.03

Table 3. Adjusted ability estimates in logits (standard errors in parentheses) for raw scores on reading tests, Forms 2B through 4B.

RAW SCORE	Form 2B	Form 3A	Form 4B
1	-5.20(1.03)	-3.83(1.01)	-2.98(1.01)
5	-3.46(.49)	-2.13(.47)	-1.29(.47)
10	-2.61(.36)	-1.39(.34)	-.49(.47)
15	-2.06(.31)	-.90(.29)	.03(.30)
20	-1.63(.28)	-.53(.26)	.42(.27)
25	-1.28(.26)	-.21(.24)	.75(.25)
30	-.97(.24)	.07(.23)	1.05(.24)
35	-.69(.23)	.32(.22)	1.32(.23)
40	-.43(.23)	.56(.22)	1.57(.22)
45	-.18(.22)	.79(.22)	1.82(.22)
50	.06(.22)	1.02(.21)	2.06(.22)
55	.29(.22)	1.25(.22)	2.30(.22)
60	.53(.22)	1.48(.22)	2.54(.22)
65	.78(.23)	1.73(.22)	2.80(.23)
70	1.04(.23)	1.99(.23)	3.06(.24)
75	1.32(.24)	2.27(.25)	3.35(.25)
80	1.64(.26)	2.59(.26)	3.68(.27)
85	2.01(.29)	2.97(.29)	4.07(.30)
90	2.50(.34)	3.47(.35)	4.57(.35)
95	3.28(.47)	4.25(.47)	5.35(.47)
99	4.94(1.01)	5.92(1.01)	7.03(1.01)
Mean	-.006	1.03	2.05
S.D.	2.32	2.22	2.30

Figure 1. Link for common item equating with approximate 95 percent confidence bands.

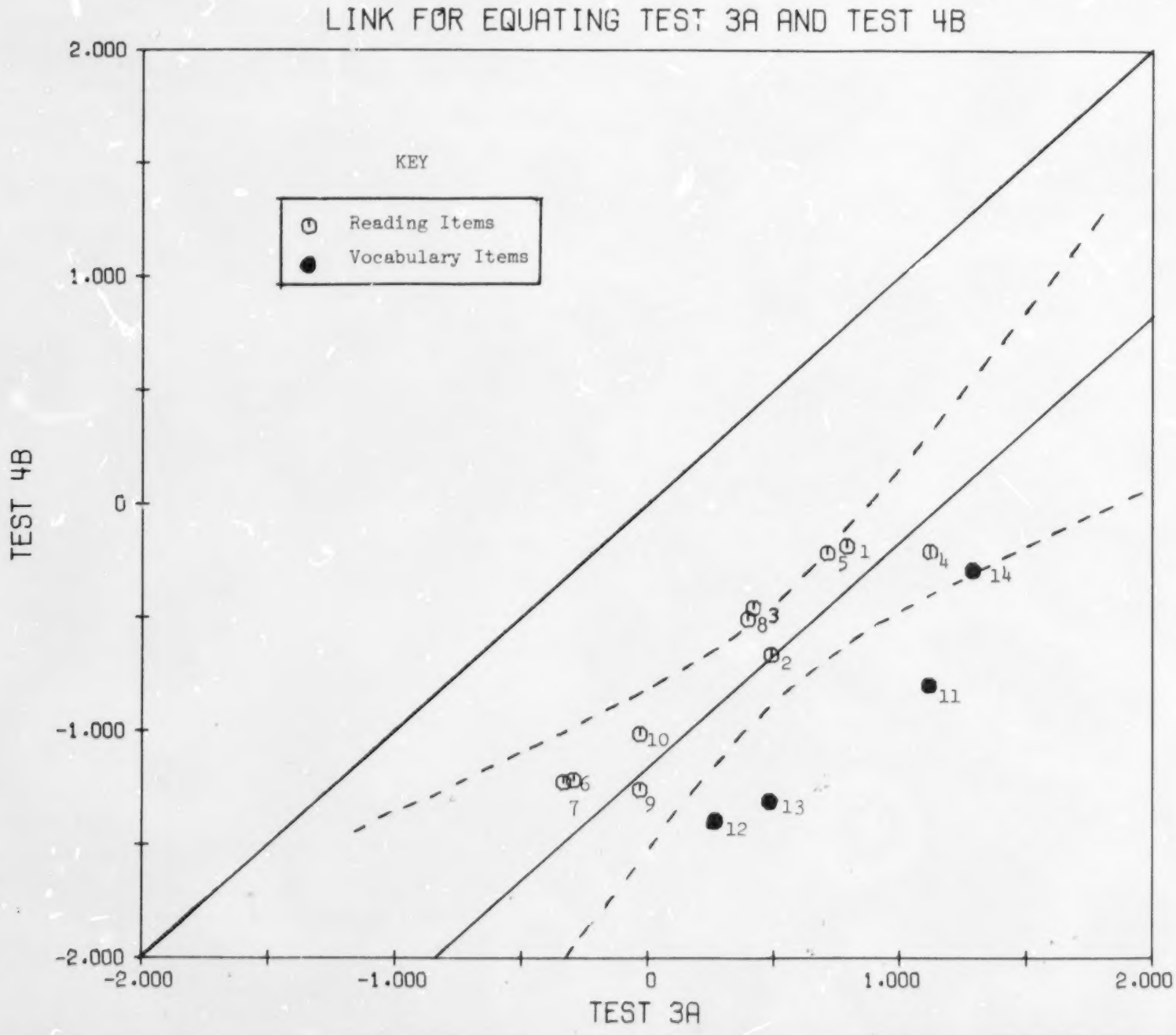
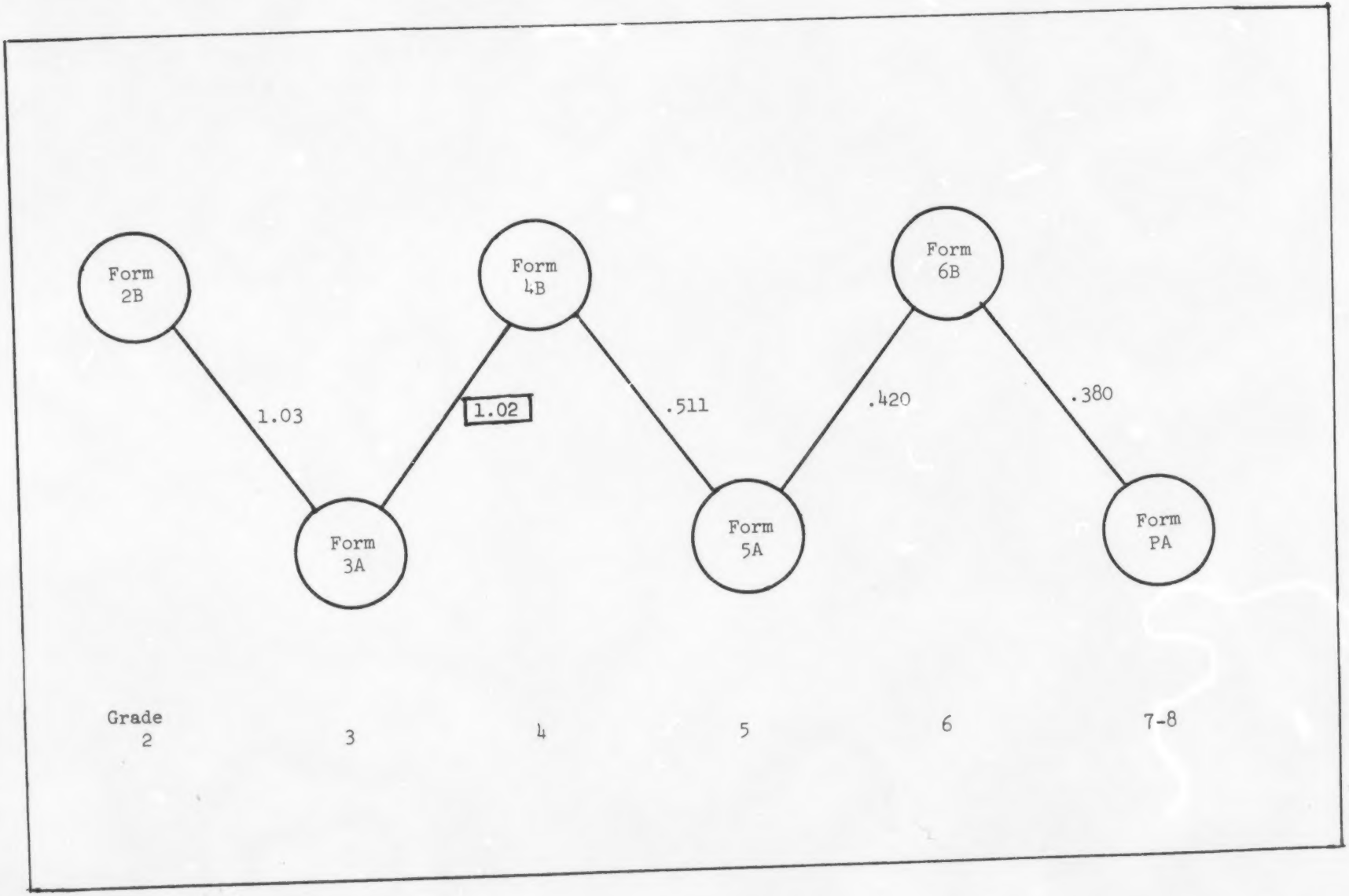


Figure 2. BA chain for reading tests, forms 2B through PA, with linking constants (Grades 2 through 7-8)



REFERENCES

Angoff, W. H. Scales, Norms and Equivalent Scores. In Thorndike, R. L. (Ed.) Educational Measurement (2nd Edition). Washington, D.C.: American Council on Education, 1971.

Engelhard, G. A Comparison of Three Methods of Scaling Psychological and Educational Tests. Unpublished manuscript, University of Chicago, 1980.

Hambleton, R. K. and Cook, L. L. Latent Trait Models and Their Use in the Analysis of Educational Test Data. Journal of Educational Measurement, 14, 2, 75 - 96, 1977.

Lord, F. M. Practical Applications of Item Characteristic Curve Theory. Journal of Educational Measurement, 14, 2, 1977, 117 - 138.

Marco, G. L. Item Characteristic Curve Solution to Three Intractable Testing Problems. Journal of Educational Measurement, 14, 2, 1977, 139 - 160.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedagogiske Institue, 1960.

Scott, Foresman and Company. Comprehensive Assessment Program.

Thorndike, E. L. On Finding Equivalent Scores in Tests of Intelligence. Journal of Applied Psychology, 6, 29 - 33, 1922.

Thurstone, L. L. A Method of Scaling Psychological and Educational Tests. Journal of Educational Psychology, 15, 433 - 451, 1925.

Thurstone, L. L. The Unit of Measurement in Educational Scales. Journal of Educational Psychology, 18, 505 - 524, 1927.

Thurstone, L. L. Scale Construction with Weighted Observations. Journal of Educational Psychology, 19, 441 - 453, 1928.

Trabue, M. R. Completion Language Scales. Contributions to Education, No . 77, Teachers College, Columbia University, 1916.

Wright, B. D. Sample-free Test Calibration and Person Measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Wright, B. D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 14, 2, 1977, 97 - 116

Wright, B. D. and Mead, R. J. BICAL: Calibrating Items with the Rasch Model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Stone, M. H. Best Test Design, Chicago: Mesa Press, 1979